Prediction of the Autism Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbor in Children

Osman Altay[1]
Department of Software Engineering, Firat University
Elazig, Turkey
oaltay@firat.edu.tr

Mustafa Ulas[2*]
Department of Software Engineering, Firat University
Elazig, Turkey
mustafaulas@firat.edu.tr

Abstract*— Autism Spectrum Disorder (ASD) negatively affects the whole life of people. The main indications of ASD are seen as lack of social interaction and communication, repetitive patterns of behavior, fixed interests and activities. It is very important that ASD is diagnosed at an early age. In this study, the classification method for ASD diagnosis was used in children aged 4-11 years. The Linear Discriminant Analysis (LDA) and The K-Nearest Neighbor (KNN) algorithms are used for classification. To test the algorithms, 30 percent of the data set was selected as test data and 70 percent as training data. As a result of the work done; In the LDA algorithm, the accuracy is 90.8%, whereas the accuracy of the KNN algorithm is 88.5%. For the LDA algorithm, sensitivity and specificity values are calculated as 0.9524 and .08667, respectively. For KNN algorithm, these values are calculated as 0.9762 and 0.80. F-measure values are calculated as 0.9091 for the LDA algorithm and 0.8913 for the KNN algorithm.

Keywords—Autism Spectrum Disorder, Autistic Spectrum Disorders in Children, Linear Discriminant Analysis, K-Nearest Neighbor

I. INTRODUCTION

ASD is thought to be a cerebral developmental disturbance that limits social communication and interaction behaviours [1]. ASD indications include many behavioral disorders, such as social interaction and lack of communication, repetitive patterns of behavior, areas of interest or activities. While ASD is seen early in development, some deficiencies and behavioral patterns may not be recognized as symptoms unless they affect the child's life in significant steps. Functional limitations vary between individuals with ASD and it is also may change over time [2].

Delay or perversion in language development is an important feature of ASD. About 50 percent of people with autism can never make a useful speech. Social communication impairment can be regarded as the most basic core feature of ASD. An example of this is in particular when babies and young children do not give the expected response to physical contact with the carer. In addition to these, repetitive behaviour and interest is a great indication of ASD. There is also unusual interest in unnecessary preoccupation, limited interests, and objects to be weirded [3].

Symptoms of ASD are usually diagnosed in children up to 18 months of age. But with this, ASD may not be recognized until the school year if the child has limited speech delay. In such cases, the diagnosis is usually made when children have problems with their friends or interactivity [4].

For diagnosis of ASD, there are various tools and approaches practiced by doctors in conjunction with diagnostic tools. In most of these studies, the classification method is used [5]. In the study conducted by Wenbo Liu et al., A new framework based on facial recognition has been proposed in addition to the use of classification methods for ASD diagnosis [6]. Another study by Yun Jiao and colleagues applied 4 different machine learning algorithms for detecting ASD. These algorithms are support vector machine (SVM), multilayer perceptron, functional trees and logistic model trees [7]. In another study, the SVM algorithm was used for ASD diagnosis [8]. In this study, sensitivity was calculated as 88% and specificity was calculated as 86%. In summary, machine learning algorithms are widely used for ASD detection.

In this study, it was tried to find out whether children have ASD by using classification methods. As a result of the classification, there are two classes of cases in which the child is ASD or not ASD. Two different classification algorithms were used. These are the KNN algorithm and the LDA algorithm. As a result of the work done, 90.8% accuracy was obtained as a result of the LDA algorithm and 88.5% accuracy was obtained from the KNN algorithm.

II. DATA ACQUISITIONS

The University of California, Irvine has a number of datasets suited to machine learning algorithms [9]. The data used in the study were obtained from a study conducted by Tabtah, F. (2017) [1]. Besides the work done by Tabtah F. has a mobile application for ASD diagnosis [10]. The data set consists of 292 samples with 19 different attributes. In the dataset, there are 10 questions directly related to ASD and a score attribute consisting of the sum of these questions.

978-1-5386-3449-3/18/$31.00 ©2018 IEEE

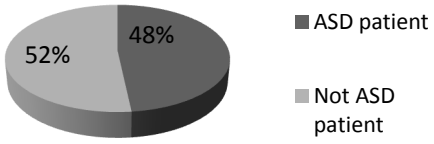
TABLE I. QUESTIONS IN DATASET

19 Different Questions	Values
It often hears voices that others do not hear.	(0-1)
It usually focuses on the big picture by going out from small details.	(0-1)
It can easily follow the conversations of different people in a social group.	(0-1)
Can easily switch between different activities	(0-1)
Does not know how to chat with her peers.	(0-1)
Good for everyday short chats.	(0-1)
It difficult for the characters to understand their intentions and feelings when he/she reading a story.	(0-1)
He/she likes to play games (role plays) that need to be imitated with other children during pre-school education.	(0-1)
He/she can easily understand what they think and feel by just looking at their faces.	(0-1)
It is difficult to make new friendships.	(0-1)
Age (4-11)	(4-11)
Gender (female, male)	(1-2)
Ethnicity('White-European', 'South Asian', 'Asian', 'Middle Eastern ', 'Pasifika', 'Hispanic', 'Turkish', 'Latino', 'Black', 'Others', 'Unknown')	(1-11)
Born with jaundice (yes, no)	(1-2)
Family member with PDD (yes, no)	(1-2)
Who is completing the test ('Parent', 'Relative', 'Health care professional', 'Self', 'Unknown')	(1-5)
Country of residence (52 different countries)	(1-52)
Used the screening app before (yes, no)	(1-2)
Scored (positive number of questions)	(0-10)

Such as ethnicity, who is completing to the test and country of residence which have a string value has been transformed to numerical values to make it suitable for LDA and KNN algorithms. All of the features included in the dataset and the 10 questions that are asked are given in detail in Table I.

The output values of the dataset consist of two classes. 19 different features have been labeled as input values and whether or not children have ASD. 1 value indicates that the child does not have ASD, whereas 2 indicates that the child has ASD. Of the 292 samples in the data set, 141 are ASD patients. The distribution of the classes in the dataset is also shown in Table II.

TABLE II. DISTRIBUTION OF DATA WITHIN CLASSES



III. ALGORITHMS AND METHODS

Two different algorithms have been chosen to predicted to ASD diagnosis with using classification method. The LDA and the KNN algorithms, which are widely used for classification. These algorithms and other methods used in the study are described below.

A. K-Nearest Neighbor

The KNN algorithm is an algorithm that gains importance by increasing the processing power of computers. Since computers have not had enough processing power before the 1960s, they are not feasible and have gained importance with the acceleration of processors. KNN algorithm is one of the most easily understood and applied machine learning algorithms in the literature. The KNN algorithm is mainly based on the distance calculation. Therefore, it is necessary to apply the KNN algorithm in numerical datasets [11]. The most commonly used method of distance calculations is the Euclidian. The equation for Euclidean calculation is given equation (1).

$$Euclidean_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \quad (1)$$

The KNN algorithm basically consists of 4 steps. In the first step, the distance from the new data to all data is calculated. In the second step, the distances are sorted. the third step takes the smallest k values and the last step determines the class.

B. Linear Discriminant Analysis

The LDA algorithm is one of the most widely used algorithms within the classification algorithms. The LDA algorithm basically operates through the calculation of variance values within and between classes [12]. In the LDA algorithm, it is necessary to calculate the scatter matrices within classes and between classes. In the LDA algorithm, it is necessary to calculate the scatter matrices within classes and

between classes. The scatter within the class is calculated as in the equation and between classes matrices are calculated as equation (2) and equation (3).

$$S_w = \sum_{i=1}^{a} \Pr(C_i)\Sigma_i \quad (2)$$

$$S_b = \sum_{i=1}^{a} \Pr(C_i)(m_i - m)(m_i - m)^T \quad (3)$$

The equation is used to calculate the Σ_i value seen in the equation (4).

$$\Sigma_i = E[(x - m_i)(x - m_i)^T] \quad (4)$$

Where x denotes the sample vector, m_i value indicates the mean value in the different classes. The equation (5) is used to calculate the discriminatory power value [13].

$$j(w) = \frac{\|w^T S_w w\|}{\|w^T S_b w\|} \quad (5)$$

The w value is the optimal discrimination projection matrix found by solving the generalized eigenvalue problem [14]. It is calculated as the result of the equation (6).

$$S_b w = \lambda_w S_w w \quad (6)$$

As a result of these calculations, the linear discriminant function is obtained by using the equation (7).

$$d(x) = w^T \left(x - \frac{m}{w}\right) \quad (7)$$

C. Performance Evaluation

In this article, commonly used evaluation measures were used to test the classification algorithms [13]. Therefore, we calculated as accuracy (8), sensitivity (9), specificity (10), precision (11) and F-measure (12) for two algorithms.

Accuracy;

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (8)$$

Sensitivity;

$$sensitivity = \frac{TP}{TP+FN} \quad (9)$$

Specificity;

$$specificity = \frac{TN}{TN+FP} \quad (10)$$

Precision;

$$precision = \frac{TP}{TP+FP} \quad (11)$$

F-measure;

$$fmeasure = 2 \times \frac{Precision \times sensitivity}{precision + sensitivity} \quad (12)$$

IV. Results and Discussion

ASD can be thought of as a disease that causes a limitation of communication and social movements as a result of deterioration in brain development. Communication ability and social behavior disorder affect people's whole life. This is why it is very important that ASD can be diagnosed early. The system for detecting ASD in children consists of 292 data samples and two classes. LDA and KNN were selected from the widely used classification algorithms for ASD detection. The Euclidean distance is used in the KNN algorithm and the k-value is taken as 1.

True positive (TP) value indicates that ASD patients are estimated by the system as ASD patients, whereas False positive (FP) values indicate that ASD is estimated as not ASD by the system. A true negative (TN) value indicates that those who do not have ASD are correctly predicted, whereas those with false negative (FN) indicate that they are misdiagnosed as not having ASD. TP, FP, TN and FN values in both algorithms are given in Table III.

TABLE III. Number of Observations

	LDA	KNN
TP	40	41
FP	2	1
TN	39	36
FN	6	9

For testing the application, 70 percent (205) of the dataset is set as training set and 30 percent (87) is set as test data. Data selection was done randomly during data sorting. Testing has been applied to both algorithms. Accuracy, sensitivity, specificity and F-measure values are also calculated for the two algorithms as a result of the test procedure. These values are shown in Table IV. The LDA algorithm has yielded relatively better results than the KNN algorithm.

TABLE IV. Evalation Measure Results

Using Algorithms	Accuracy	Sensitivity	Specificity	Precision	F-Measure
LDA	0.9080	0.9524	0.8667	0.8696	0.9091
KNN	0.8851	0.9762	0.8000	0.8200	0.8913

For the LDA algorithm, the sensitivity value was calculated as 0.9524, the specificity value was calculated as 0.8667, the precision value was calculated as 0.8696, while the F-measure value was calculated as 0.9091. In the LDA algorithm, the accuracy value was calculated as 0.9080. In the KNN algorithm, the accuracy value was calculated as 0.8851, while the sensitivity value was calculated as 0.9762, the specificity value was calculated as 0.80, the precision value was calculated as 0.82, and the F-measure value was calculated as 0.8913.

V. Conclusions

In this study, children aged 4 to 11 years were tried to be diagnosed by ASD disease classification method. A dataset with a wide range of questions was used in the study. The LDA and the KNN algorithms are used in the classification methods. There is no operation on the data except to convert it to numerical values. Accuracy, sensitivity, specificity, precision and F-measure values are calculated and the values

of TP, TN, FP and NP are given in order to make a comparison with the next studies.

As a result of the implementation, the LDA algorithm gave a better result than the KNN algorithm at the accuracy value. But the KNN algorithm is more successful than LDA in the sensitivity value, which is a value indicating that children patients are ASD patients. The F-measure value is calculated as 0.9091 for the LDA algorithm and 0.8913 for the KNN algorithm. The LDA algorithm in the F-measure value provided a 1.95% better success rate than the KNN algorithm.

REFERENCES

[1] Thabtah, F. (2017, May). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. In Proceedings of the 1st International Conference on Medical and Health Informatics 2017 (pp. 1-6). ACM.

[2] Autism and Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators. (2014). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2010. Morbidity and Mortality Weekly Report: Surveillance Summaries, 63(2), 1-21.

[3] Haq, I., & Le Couteur, A. (2004). Autism spectrum disorder. Medicine, 32(8), 61-63.

[4] Blumberg, S. J., Bramlett, M. D., Kogan, M. D., Schieve, L. A., Jones, J. R., & Lu, M. C. (2013). Changes in prevalence of parent-reported autism spectrum disorder in school-aged US children: 2007 to 2011-2012 (No. 65). US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

[5] Maenner, M. J., Yeargin-Allsopp, M., Braun, K. V. N., Christensen, D. L., & Schieve, L. A. (2016). Development of a machine learning algorithm for the surveillance of autism spectrum disorder. PloS one, 11(12), e0168224.

[6] Liu, W., Li, M., & Yi, L. (2016). Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. Autism Research, 9(8), 888-898.

[7] Jiao, Y., Chen, R., Ke, X., Chu, K., Lu, Z., & Herskovits, E. H. (2010). Predictive models of autism spectrum disorder based on brain regional cortical thickness. Neuroimage, 50(2), 589-599.ISO 690

[8] Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., ... & MRC AIMS Consortium. (2010). Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage, 49(1), 44-56.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[9] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

[11] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18.

[12] Zhao, W., Chellappa, R., & Nandhakumar, N. (1998, June). Empirical performance analysis of linear discriminant classifiers. In Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on (pp. 164-169). IEEE.

[13] Jain, A., & Huang, J. (2004, May). Integrating independent components and linear discriminant analysis for gender classification. In Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on (pp. 159-163). IEEE.

[14] Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), 3358-3378 (2007).